Proceedings of the 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 14-17 July, 2013

MULTI-PATTERN FUSION BASED SEMI-SUPERVISED NAME ENTITY RECOGNITION

ZIGUANG CHENG, DEQUAN ZHENG, SHENG LI

MOE-MS Key Laboratory of Natural Language Processing and Speech Harbin Institute of Technology, Harbin, China 150001
E-MAIL: {zgcheng, dqheng, lisheng}@mtlab.hit.edu.cn

Abstract:

Named Entity Recognition (NER) is one of the most important problems in Natural Language Processing (NLP). NER also has a broad prospect for application and important research value. There are a lot of methods and technology to solve NER problem. In this paper, for a specific application background, a new multi-pattern fusion based semi-supervised NER method is proposed. We use soft-matching method in entity internal pattern first. Then through bootstrapping process in the training corpus, we get an entity external pattern. Finally we use fusion internal and external pattern method to complete the named entity recognition. Experiments on Chinese weapon names, from People's Daily corpus and some military news articles were performed. They showed when the internal characteristic is significant and training corpus has a higher similarity with test corpus, this method performs better than soft matching method and external pattern based bootstrapping method, improving the named entity recognition precision by 18.2%.

Keywords:

Name entity recognition; soft matching; bootstrapping; multi-pattern fusion

1. Introduction

The term "Named Entity (NE)", is widely used in Information Extraction (IE), Question Answering (QA) and NLP applications. During the course of system development, people noticed that it is important to recognize information units like names (person, organization and location names) and numeric expressions such as: time, date, money and percent expressions[1].The expansion of IE applications gave rise to the need for a drastic extension of NE categories. The extension of NE categories brings difficulty to NER[2]. Now, the number of categories is limited to 7 to 10. And the NE taggers, automatic annotation systems for NE entities in unstructured text, are based on 1) dictionaries and rules which were made by hand, or 2) some supervised learning techniques. More recently and currently, dominating technology is the supervised learning techniques. Since it is quite difficult to obtain large scale and high quality NE-annotated corpora, the supervised-learning methods used for traditional NER are not easily adaptable to new categories[8]. But, bootstrapping, a semi-supervised learning method, is attractive to get patterns from unlabeled corpus for extended NER, because it only needs a small set of seeds and a small degree of supervision[3]. Bootstrapping and pattern-based methods are widely used in information extraction (IE), in which patterns are used to start extraction and extract the relevant information to a particular task[4]. Besides that, this method can also be used in NER and has a not bad result. However, its precision and recall are still not satisfying.

The main reason is that only using internal pattern or external pattern often has limited coverage since their structures are fixed and include less information[5-8]. Another probable reason lies in the exact pattern match, i.e. all the parts of a pattern must be matched[9],[10].

This paper focuses on fusion internal and external pattern, using soft-matching[11] and bootstrapping[6] technology to improve NER performance.

2. Soft-matching and bootstrapping

To get NE external pattern, we use soft matching and NE internal pattern through bootstrapping process in training corpus. In this section and next section, we will describe soft matching method, bootstrapping especially our fusion method in detail.

2.1. Soft matching

Though the internal compositions of NE varies from each other, there are certain regularity underlying these compositions. Take Chinese weapon name as an example: weapon "歼-10 战斗机" (jian-10 fighter), its internal composition is: type name (歼) + delimiter (-) + type name (10) + weapon headword (战斗机). Take existed weapon

978-1-4799-0260-6/13/$31.00 ©2013 IEEE

names as internal pattern, then use soft matching algorithm to calculate the similarity of the $entity_{option}$ (entity to be matched) and every pattern. If the result is high enough, the entity will be recognized as an NE.

To calculate the similarity more precise, this article adopts Levenshtein Distance Method to calculate the distance between two patterns. If similarity reaches a pre-set threshold this entity will be considered to be matched. Levenshtein Distance is used for conversing one mode into another to add, delete and modify the value.

Suppose there are two modes $X(x_0,x_1,..,x_i)$ and $Y(y_0,y_1,...,y_j)$ Levenstein Distance could find the minimum cost to transform from mode $X(x_0,x_1,...,x_i)$ to mode $Y(y_0,y_1,...,y_j)$. The Levenstein Distance is defining as formula (1).

$$Levenshtein(x_i,y_j)=\begin{cases}0 & x_i=y_j\\ p & x_i\neq y_j\\ q & insert\ y_j\end{cases} \quad (1)$$

Formula (1) is Levenshtein distance consideration formula, x_i is an element of the mode X , y_j is an element of the mode Y , p and q are pre-set constants.

(1) If x_i is equal to y_j , then x_i (or y_j) won't affect the distance between X and Y .

(2) If x_i is not equal to y_j , substitution will be performed, and the distance between X and Y will increase p.

(3) If x_i is not equal to y_j and insert y_j into X , X and Y will be matched, then the distance between X and Y will increase q.

With Formula (1), we can calculate the distances between $entity_{option}$ and every internal pattern, and the minimum distance (entity identifiable degree) will decide whether the entity can be matched. The calculation process of this entity identifiable degree is shown in Algorithm 1.

Algorithm 1 pattern similarity calculation

1. If the length of X or Y is zero, return the length of the other pattern;
2. Initialize (i+1)*(j+1) matrix D, the first row and column set to zero, i is the length of X , j is the length of Y ;
3. scan this two pattern (i*j degree), if $x_i = y_j$, set variable temp to 0 otherwise set temp to p, then set D[i][j] to the minimum of d[i-1][j]+q, d[i][j-1]+q, d[i-1][j-1]+temp.
4. return D[i][j] as the distance between pattern X and pattern Y .

end

2.2. NE hard pattern obtainment

Bootstrapping is the process of improving the performance of a trained classifier which is often used in situations where labeled training data is scarce but unlabeled data is abundant [6]. For NER tasks we also have a lot of training data, but less labeled data.

In this paper, together with soft matching, the bootstrapping method mainly use already exist named entity internal pattern, to get named entity external patterns.

2.2.1 External pattern obtaining

NER external pattern is mainly composed of the extracted items and feature items. Extracted item indicates the NE to be extracted; the feature item is used to determine the context information of extracted items to play the role of the trigger and constraint checking.

The specific form of the external pattern is shown as follow[7]:

<Token_L>...<Token_2><Token_1> INTEREST_CLASS < Token_+1>< Token_+2> ... < Token_+L>.

L is the length of external pattern.Slot(< *Token_i*>)can be a word form, POS (part of speech) or semantic category information. In most cases, the slot contains a word form. We substitute numbers, letters, punctuation, time and job titles with their POS or semantic category. Specific steps bootstrapping and soft matching fusion for obtain external pattern are as follows:

(1) Manually initialize seed set: S_{seed} .

(2) Extract external patterns with seed set, and add new patterns to the established candidate pattern set: P_{cand} .

(3) Add accepted external patterns to the available pattern set: $P_{accepted}$.

(4) With seed set and $P_{accepted}$, we use soft matching method to identify NEs, establish candidate instance set.

(5) If there are no new named entity being recognized, or reaching a certain number of iterations, then loop ends; otherwise go to (6).

(6) According to credible external pattern evaluation of candidate instance, determine the new seeds and return to step (2).

The external pattern acquisition process is shown in Figure 1.

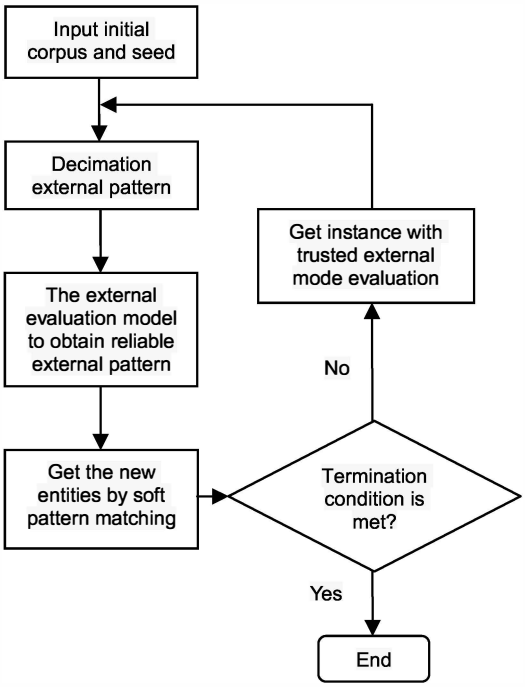

Figure 1.Chart of external pattern acquisition based on bootstrapping

2.2.2 Pattern evaluation

In external pattern obtain processing, instance and pattern depend on each other. Reliable instance set constructs reliable instance set, then reliable pattern set construct reliable pattern set. So evaluating pattern and instance is very necessary. Formula (2) and Formula (3) are used to evaluate patterns and NE instances.

$$Score(P_j) = 1 - \frac{Num_{CommWord}(P_j)}{TotalNum_{Word}(P_j)} \quad (2)$$

$$Score(NE_j) = \frac{1}{n}\sum_{i=1}^{n} Score(P_i) \quad (3)$$

In Formula (2), $Num_{CommWord}(P_j)$ is the number of the common words discovered by the pattern P_j, $Total Num_{Word}(P_j)$ is the total number of words extracted by the pattern P_j. Common words are the words appearing in a basic dictionary. In Formula (3), P_i is one of the patterns, which can extract NE_j in the current iteration is the total number of this kind of patterns.

3. Internal and external pattern fusion NER

The main process of multi-pattern fusion NER is: (1)Use Soft matching and bootstrapping fusion method get external pattern; (2) Get external soft pattern based on external pattern set and weight information; (3) Extract the context information of the $entity_{option}$; (4) Use soft matching method to calculate internal pattern matching degree; (5) Calculate the external pattern matching degree; (6) Jointly consider internal pattern matching degree and external pattern matching degree recognized NE; (7) With certain rules resolve conflicts.

The overall process of multi-pattern fusion based semi-supervised NER is shown in Figure 2.

3.1. Generalizing external hard patterns to soft patterns

The information in every slot of external soft pattern could get from the external hard pattern, which is obtained in the previous section. The key is how to define and obtain weight information.

We will take each token in the occurrence probability of each slot as the corresponding weight. Specifically, we calculate the weight according to the Formula (4) .

$$W_{l,i} = P(Token_{l,i}) = \frac{Num(Token_{l,i})}{TotalNum(Token\_in\_slot_l)} \quad (4)$$

Wherein, $Num(Token_{l,i})$ is the number of occurrences of the $Token_i$ in the $slot_l$. These parameters can be obtained through the hard pattern set. The following is a specific example of a soft pattern. Specifically, we will use the logarithm opposite number of the weighted information. Note: every slot in the soft pattern is a BOW.

<slot-3><slot-2><slot-1> INTEREST_CLASS <slot+1><slot+2><slot+3>

START 2.4 ; 据 6.8 ; 新华社 6.7; interest_Class; TTIME 4.6; TTIME 2.7 ; 电 6.3

被 6.8; 推选 10.4 ; 为 4.8 ; interest_Class; 委员 4.7; PUNC 2.6 ; END 2.1

In this example, START and END respectively represent the start and end of a sentence. Soft pattern matching not simply the two values match. Instead, for each test word or phrase, first, extract the corresponding context; then calculate the context score with the weight information of each slot information in the external soft pattern. If the score threshold condition is satisfied, it is considered found interesting information. Specifically, use the Formula (5) and Formula (6)

calculates the score of the context to be measured.

$$Score(Pattern) = \prod_{1} P(Token_{l,i}) \quad (5)$$

$$Seq\_Score(Pattern) = \prod_{1} P(Token_{l+1,j} | Token_{l,i}) = \prod_{1} \frac{Num(Token_{l,i}, Token_{l+1,j})}{Token_{l,i}} \quad (6)$$

Formula (5) is used to calculate the product of the probability values of each slot. $P(Token_{l,i})$ can get from Formula (4). Now suppose the emergence of each slot is independent, the formula (5) is the joint probability of each slot. It measures how well context information includes the possibility of a specific semantic category. Formula (6) is used of bigram to calculate, calculate the probability of the slot sequence. Wherein, $Num(Token_{l,i}, Token_{l+1,j})$ is the number of the co-occurrence of $Token_i$ in $Slot_l$ and $Token_j$ in $Slot_{l+1}$.

With the Equation (5) and Equation (6), the degree of pattern matching can be calculated. If these two values are greater than the pre-set threshold, we will consider that context to be tested contains interested entities. In the recognition process, some parameters may not appear in the training corpus, so we need parameter smoothing to ensure that the evaluation pattern matching. We will take the minimum value of corresponding parameters in the corpus as the current values of the parameters in the calculation.

3.2. Multi-pattern fusion NER method

For the internal pattern distinct named entity, soft matching has a good effect. Take weapon NE for example: if we already have NE “BGM-109‘战斧’巡航导弹” ("BGN-109 'zhanfu' cruise missile"), we could recognize “BGM－109 巡航导弹” through soft matching method; or we already have NE “歼-10 战斗机” (jian-10 fighter), then through soft matching method we could recognize “歼-20 战斗机” (jian-20 fighter)、“歼十战斗机” (jian ten fighter) or some other similar weapon NE. But soft matching method cannot resolve the conflict well. So only by using internal pattern the desired requirements cannot achieved.

The process of multi-pattern fusion NER method is as follow. At first we treat a possible word series, which based on a rough segmentation, as a $entity_{option}$. Then we get its corresponding internal and external features, and use Formula (5) and Formula (6) to calculate the external score At least we calculate the $entity_{option}$ internal score with existing internal pattern though soft matching. So far we have obtained the internal and external pattern score of $entity_{option}$. Next we will decide whether it is recognized as an NE. If the internal score is over the fixed internal pattern threshold and the external score is over threshold T_1 (Because we will used a set of external pattern threshold later, this group threshold is defined as T_1) we will put the $entity_{option}$ to the candidate result set. If the internal scores lower than the fixed internal pattern threshold and the external score is over threshold T_2, we also put the $entity_{option}$ to the candidate result set. Obviously the threshold T_2 is strict than threshold T_1. Finally we will use certain rules to resolve the conflict for every entity in candidate result set. The overall process of multi-pattern fusion NER is shown in Figure 3.

In order to ensure the system's recall rate, we have taken all possible word series as $entity_{option}$. If the $entity_{option}$ satisfy the rules in the sectionabove, then the $entity_{option}$ will be put into candidate result set. So we need to resolve conflicts. Take the following rules to solve the conflict:

(1) Satisfy both internal and external patterns of priority. Simultaneously met in the case, the $entity_{option}$ whose external score is highest will as the final result. The external score consider the rule (2) to (3).

(2) The entity with bigger joint probability has priority.

(3) The entity with biggest length has priority

4. Experiment and analysis

We tagged the weapon names in the People’s Daily newspaper (1988.01-1988.02) and extracted the corresponding paragraphs to form the testing set. The training corpus is the People’s Daily corpus (1988.03-0988.06), among which the corpus of Jun is used as validation set to determine the best threshold. The initial seeds are manually selected and not more than 20. And the soft matching method needs a domain dictionary. An entity is regarded to be correctly recognized if the boundary and category are both correct.

According to the feature of the weapon NE and the characteristics of the soft matching method, divided NE into two parts according to weapon NE's internal feature. We calculate the experimental results of these two parts respectively. The first category, internal pattern distinct weapon NE, such as:“‘复仇者’防空导弹系统” (" 'avenger' air defense missile system")and“‘G 级’弹道导弹常规潜艇” (" 'G degree' Conventional Submarine Ballistic Missile"). The

other category , common weapon NE, such as: “子弹” (bullet)，“枪支” (firearms)，“火箭” (rocket). The NER experiments will be carried out with soft matching method, external pattern based bootstrapping and fusion method. The experimental results of the two parts of the NE statistics are shown in Table 1, Table 2 .

TABLE 1. **INTERNAL PATTERN DISTINCT NE EXPERIMENTAL RESULT TABLE**

	Recall	Precision	F
soft matching	74.2	45.6	56.4
bootstrapping	51.6	27.8	36.1
fusion method	78.6	63.8	70.4

As we can see in Table 1, with regard to the internal pattern distinct weapon NE. Soft matching method is better than external pattern based bootstrapping method. However, when applied softmatching method to distinct internal name entity, the recall rate is relatively high, but on the contrary, the accuracy is quite lower. The experiment also showed that the main inaccurate boundary identification, so the fusion method can effectively improve the precision and a slight increase recall.

TABLE 2. **COMMON WEAPON NE EXPERIMENTAL RESULT TABLE**

	Recall	Precision	F
soft matching	28.1	18.5	22.3
bootstrapping	55.3	35.3	43.1
fusion method	72.8	50.1	59.4

As we can see in Table 2, with regard to the common weapon NE. When applying soft matching method to common weapon NE, both the precision and recall is very low, and the external pattern based bootstrapping method the effect is not good. But fusion method could effectively improve the recall and precision.

4. Conclusions

In this article we propose a multi-pattern fusion based semi-supervised NER method. Experiments on People's Daily corpus and 100 military news, shows that compared with the method before fusion, fusion method has a higher precision and recall.

In future work, we will try to merge other pattern matching method to improve the NER performance. And we will increase the scale of experiment.

Acknowledgements

This work is supported by the national natural science foundation of China (No. 61073130) and the project of National High Technology Research and Development Program of China (863 Program) (No. 2011AA01A207).

References

[1] D. Nadeau, S. Sekine, “A survey of named entity recognition and classification,” Linqvisticae Investigationes, Vol 30, Number 1,2007, pp. 3-26(24)

[2] Satoshi Sekine, Named Entity: History and Future, 2004, available on-line at http://cs. nyu. edu/sekine/papers/

[3] M.Thelen, E.Riloff. “A bootstrapping method for learning semantic lexicons using extraction pattern contexts”, Proc. of the 2002 conference on Empirical Methods in Natural Language Processing(EMNLP), Philadelphia, USA,pp.214-221,2002.

[4] W.Lin, R.Yangarber, R.Grishman, “Bootstrapped Learning of Semantic Classes from Positive and Negative Examples”, Proc. of ICML2003 Workshop on the Continuum from Labeled to Unlabeled Data, Washingtong DC, USA, pp.103-101, 2003.

[5] Tan Hongye, Zhao Tiejun, Yao Jianmin. “A Study on Pattern Generalization in Extended Named Entity Recognition”. Chinese Journal of Electronic, 2007, 16(4): 675-678.

[6] Hang Cui, Min-Yen Kan, Tat-Seng Chua,“Unsupervised Learning of Soft Patterns for Generating Definitions from Online News”, Proceedings of the 13th World Wide Web Conference, New York, USA, pp.90-99, 2004

[7] R. Yangarber, W. Lin, R. Grishman, “Unsupervised Learning of Generalized Names—Proceedings of the 19th International Conference on Computational Linguistics”, Taipei, Taiwan, 2002. pp. 1135–1141.

[8] Jiang Xiao, Tat-Seng Chua, Huang Cui, “Cascading Use of Soft and Hard Matching Pattern Rules for Weakly Supervised Information Extraction” COLING '04 Proceedings of the 20th international conference on Computational Linguistics Article No. 542

[9] Un Yong Nahm, Raymond J.Mooney, “Mining Soft-Matching Association Rules” CIKM '02 Proceedings of the eleventh international conference on Information and knowledge managementPages 681-683 ACM New York, NY, USA 2002

[10] Yu Chen, Dequan Zheng, Bowen Zheng, Tiejun Zhao, “Research on Automatic Pattern Acquisition Based on Construction Extension”, JCIT 01/2010; 5:122-127

[11] Chen Niu, Wei Li, Jihong Ding, Rohini K.Srihari, “A bootstrapping approach to named entity classification using successive learners”, Proceedings of the 41st Annual Metting on Association for Computational Linguistics, Sapporo, Japan (2003), pp.335-342

[12] Yusuke Shinyama, Satoshi Sekine, “Named entity discovery using comparable news articles”. InProc. The International Conference on Computational Linguistics (COLING),2004,pp.848-853.

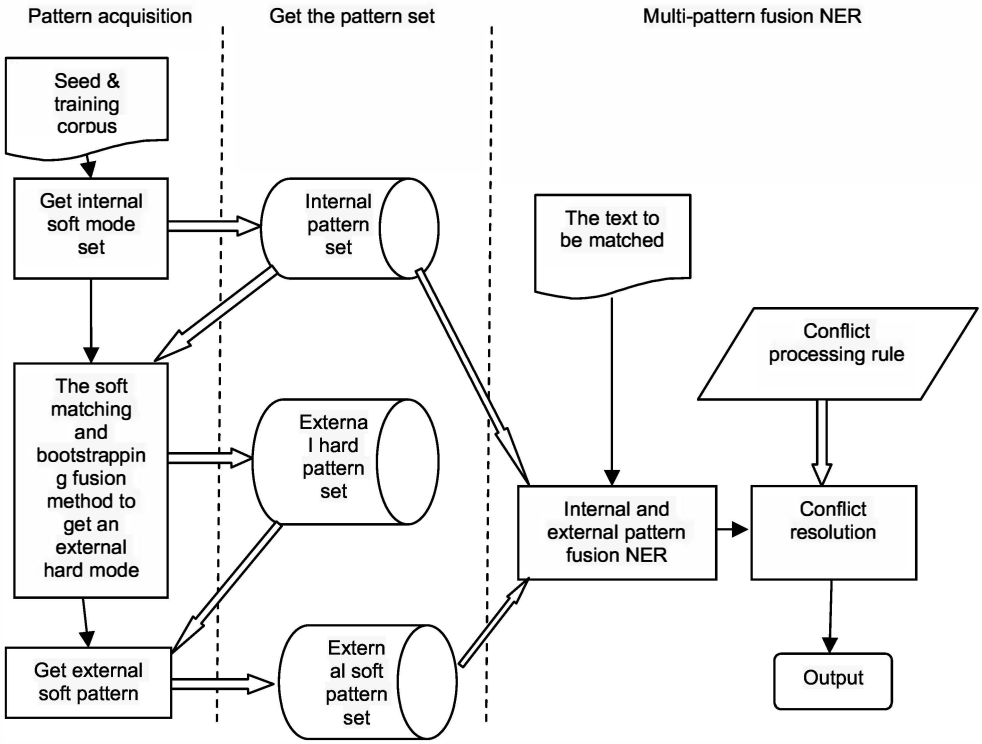

Figure 2. Based on overall schematic diagram of the multi-pattern fusion method.

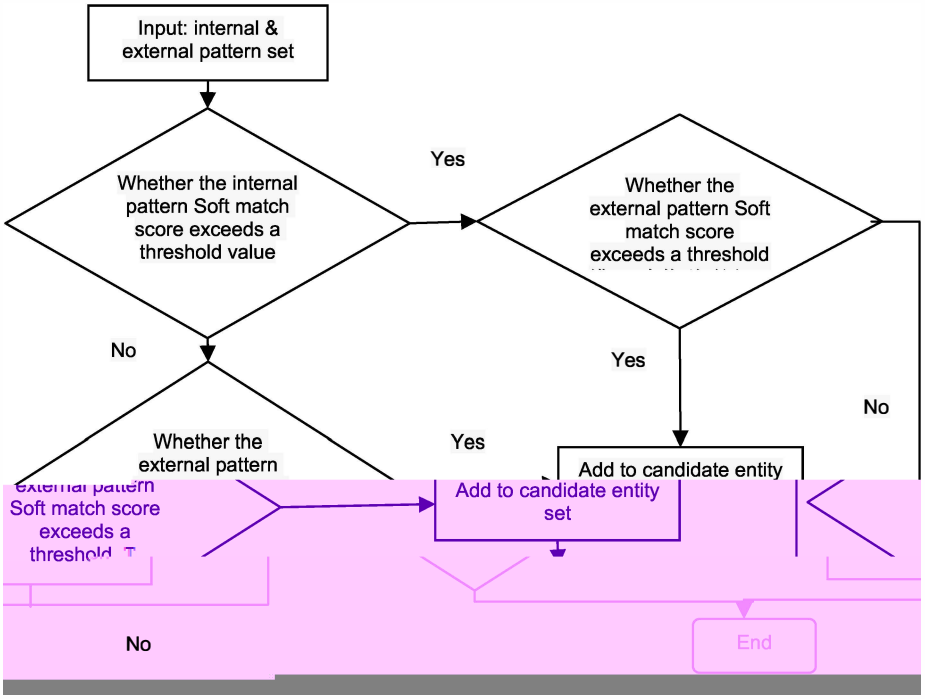

Figure 3. Schematic diagram of multi-patternNER method.